Mr. Robert Graybill
Information Processing Technology Office (IPTO)
High Productivity Computing Systems

Good morning!

As an introduction, I would like to play a short video clip highlighting a major new IPTO initiative in high productivity computing systems.

Video Clip (3.5 minutes)

As you have just observed, high performance computing is at a critical juncture. Over the past 3 decades, this important technology area has provided vital computational capability for many important national security applications. Government research, including substantial DoD investments, has enabled major advances in computational capabilities, contributing to U.S. dominance of the world computer market. Unfortunately, current evolutionary trends in commercial high performance computing, emerging computer architecture technology barriers, and emerging threats are creating computing capability gaps that threaten continued U.S. superiority in important national security applications.

Today's high-end systems tend to fall into one of two domains: the vector supercomputer domain or the commodity high performance computer domain. Foreign computer vendors dominate the vector domain with Cray as the sole domestic supplier. A majority of the terascale computing installations in the United States consist of commodity HPCs.

The High Productivity Computing Systems Program was initiated to address the very real and expanding HPC capability gap documented by a number of DoD studies. The HPCS initiative incorporates strong collaboration and sponsorship from a number of key Government agencies such as NSA, NRO, DOE, NNSA, and DDR&E. The High Productivity Computing Systems Program will bridge the gap between the late-80s-based technology of today's high performance computers and the promise of quantum computing for the Department of Defense.

DARPA's challenge is to develop a broad spectrum of innovative technologies and architectures integrated into a balanced total system solution by the end of this decade. I want to highlight a subtle, but important, change in emphasis from past programs in high-end computing. The new emphasis is now on productivity or value.

Using raw theoretical peak computing performance as the single evaluation criteria is not sufficient. Total end-user computing life-cycle costs and mission responsiveness are critical to future large tera- and peta-scale computing installations and end users. An analyst's idea-to-solution or time-to-solution is more important than raw computing capacity. It is time we move beyond Moore's Law, the doubling of microprocessor performance every 18 months, and double the productivity value instead.

The end product of the HPCS Program will be economically viable, high-productivity computing systems with both scalable vector and commodity functionality for national security and industrial user communities. These objectives cannot be met simply by tracking Moore's Law and leveraging evolutionary commercial developments, but will require a revolutionary technology step function increase in computing capability. Without Government-sponsored R&D and participation, the only available high-end computing solutions will be those with a strong mass-market consumer and business bias. For example, grid computing systems, consisting of a large number of loosely coupled computers, will not provide the high-bandwidth low-latency computing required for the DoD applications I will describe next.

The HPCS Program will restore information superiority to the soldier in a number of key application areas. In the near term, these include operational weather and ocean forecasting, cryptanalysis, and planning exercises related to dispersion analysis of airborne contaminants. In the longer term, these critical areas are weapon design such as warheads and penetrators; survivability and stealth design; intelligence, surveillance

and reconnaissance systems; virtual manufacturing and failure analysis of large aircraft, ships and physical structures; and emerging biotechnology.

A study conducted in April 2001 found the results of operational weather forecasting codes run on commodity HPCs available to DoD put the U.S. at a 5-year disadvantage to European nations.

In the area of cryptanalysis, the challenges encountered in using commodity HPCs may lead to the dilution of research and operational readiness. Rapid crises response to the dispersion of airborne contaminants will be enhanced by providing increased computational modeling capability and resolution. The HPCS Program will create and supply new systems and software tools that will lead to improved time-to-solution and computational capabilities. These will result in significant operational capability advances and responsiveness across a broad spectrum of national security areas.

Four major HPCS goals have been established to support the new emphasis on high end computing productivity or value.

- *Performance* aims to improve computational efficiency and performance of critical national security applications 10 to 40 times over today's scalable vector and commodity high performance solutions. Currently, computational efficiencies are as low as 5 percent for some applications running on terascale installations.

- *Productivity* endeavors to reduce the cost of developing, operating, and maintaining high-end computing applications. The total cost of ownership far exceeds the original cost of a terascale computing facility. As an example, recent data from a telecom company shows that its server and workstation maintenance costs will exceed $1.9 billion per year.

- *Portability* seeks to insulate research and operational HPC application software programmers from system architecture details without decreasing overall system efficiency. Example architecture details are processor type and quantity, memory organization, and communication models. A recent study indicates that it is 100 times more difficult to program a large parallel machine than a single processor. Software development productivity has improved by a factor of only 10 over the last 30 years for non-HPC class computers.

- *Robustness* strives to deliver improved reliability to HPCS users, through both conventional and innovative means. The more traditional forms address inherent system brittleness and susceptibility of large scale systems to conventional hardware and software reliability issues. In today's environment, robustness also requires addressing unconventional reliability issues such as recovering from logic soft errors, software bug tolerance and intrusion resistance.

To achieve this aggressive goal of revolutionary HPCS solutions by the end of this decade, three top-level program phases have been identified to address the challenges of scalable vector and commodity HPC solutions: concept study, research and development, and full-scale development. The 1-year Phase 1 industry concept study initiated this year will provide critical technology assessments, develop revolutionary HPCS concept solutions, and supply new productivity metrics necessary to develop a new class of high-end computers by the end of this decade. The results from Phase 1 will be merged with other DoD long-range high-end computing mission requirement assessments to form the basis for Phase 2.

I am very happy with the response from the industrial and research communities to the Phase 1 broad area announcement released early this year. Phase 1 participants have been selected and are now under contract. Industry HPCS concept awards have been made to industry teams led by Cray, HP, IBM, SGI, and Sun. These teams also include more than 20 universities and research organizations.

A major challenge is formulating a comprehensive set of requirements, scalable benchmark strategies, and metrics for these tera- and peta-scale computing systems. Technical input from industry, Government laboratories, and DoD end users will be encouraged. Mitre Corporation is leading the applications analysis and performance assessment team.

The second phase of the HPCS program is a 3-year research and development effort that will perform focused R&D and risk-reduction engineering activities. These pursuits will result in a series of system design reviews, preliminary design reviews, and risk-reduction prototypes and demonstrations. The technical challenges and promising solutions identified during the concept study will be explored and prototyped by a full complement of commercial industry, university, and research laboratory researchers.

Phase 3, full-scale development, will be led by commercial industry. This phase will last 4 years and complete the detailed design, fabrication, integration and demonstration of the full-scale HPCS pilots SN001. The goal is to provide early hardware and software releases or scaled-down replicas of the final pilot systems for early evaluation by universities and DoD end users. The final pilot systems will be delivered to selected critical DoD end users. The current plan is to encourage the research community to evaluate HPCS research platforms and software development environments and to explore long-term, high-end computing challenges.

To meet these technical challenges, we must develop a balanced total solution incorporating end-user requirements. DoD operational and research software applications will serve as the requirements driver for architecture and software research and systems assessment. The key is developing meaningful scalable productivity value metrics.

Just what are value-based metrics? Examples include:

- Robustness
- Idea-to-solution
- Time-to-solution
- Application life-cycle costs
- Ownership costs -- facilities, support staff, training
- Acquisition costs -- infrastructure and equipment
- System scalability
- Evolvability from flops to petaflops.

The top-level technical challenge will be developing concepts for a productive system with the ability to scale with technology and double in value every 18 months over the next 2 decades. The solution will be to identify and assess critical technologies and leverage research efforts being performed by universities, research laboratories, and Government agencies.

Five basic technical focus areas have been identified: high effective bandwidth with low latency, balanced system architecture, system tailorability, robustness, and performance measurement and prediction. The HPCS Program will address these technical component areas at all levels in the system: programming models, hardware, software, system architecture, and performance monitoring.

These five technical challenges are not fully addressed by today's HPC system architectures, which have a commodity market orientation. In general, they are relatively unbalanced from a software and hardware perspective as they need to support a broad spectrum of customers, programming methodologies, and standards. As a result, we are observing that some complex DoD-derived simulation codes demonstrate poor computational efficiency. This is a real technical barrier because the physical sizes of our current teraflop systems are starting to exceed our capacity to house and maintain them. A fundamental issue is the imbalance among processor speed, system latency, and bandwidth. For complex simulations, this imbalance can result in application software with poor scalability across large numbers of processing nodes resulting in single processor performance, as low as 5 percent of peak processor performance

New innovative programming models and architectures must be developed to investigate and implement software tools for program production and understanding. For example, program understanding tools would allow programmers and users alike rapid insight into both the correctness and performance of their application software. A very promising research area is the ability to perform online profiling correctness to

support software bug tolerance and intrusion resistance capability. Finally, enhancements to operating and runtime systems will be required to support these objectives.

The responsive programming models, architectures, and profiling techniques I have described will require multilevel adaptation. This may mean either software adapting to a given hardware environment or system hardware adapting to a given software load. Current efforts in performance monitoring and software adaptability are beginning to pay dividends by giving us the tools to choose good implementations of algorithms for targeted architecture and available computing resources automatically.

Examples of early research in responsive micro-architectures are DARPA's Data Intensive Systems, Power Aware Computing and Communications, and Polymorphous Computing Architectures programs. HPCS will leverage this early research and extend it by creating fully responsive high-end systems that mutually seek the best operating point for a given problem or mission area.

The technical ideas, major goals, and program objectives outlined in this talk were derived from a broad segment of industry, research, and DoD user communities over the last 2 years. Speaking for these communities, we now have the opportunity we have been waiting for: to provide a revolutionary step function increase in productivity in high end computing. Are we up to it? I believe we are! Aside from the technical aspects, stable multi-agency technical management support and full funding for this critical, long-term program over the balance of this decade are mandatory to close our nation's high-end petaflop computing national security gap!

We encourage the continued active participation of all of you warfighters, Defense contractors, commercial industry, and the research community in this exciting, major, long-term initiative.

I am looking forward to working with you in the future.

Thank you for your time!